Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
53 t., 2012, 90–95

ISSN 0132-2818
www.mii.lt/LMR/

procedūra trumpų laiko eilučių prognozei

Rita Palivonaitė, Minvydas Ragulskis

Kauno technologijos universitetas, Fundamentaliųjų mokslų fakultetą
Studentų g. 50-325, LT-51368 Kaunas

E. paštas: rita.palivonaite@ktu.lt, minvydas.ragulskis@ktu.lt

Santrauka. Straipsnyje pateiktas patobulintas trumpų laiko eilučių prognozės metodas, identifikuojantis algebrinės sekos skeletinę kreivę. Prognozei pagerinti naudojama vidinio glodinimo procedūra. Eksperimentai atlikti su dirbtinai sugeneruota testine eilute ir realiais duomenimis.

Raktiniai žodžiai: Hankelio matrica, laiko eilučių prognozė, algebrinė seka.

Įvadas

Laiko eilučių prognozavimo uždavinys yra aktualus daugelyje mokslo ir inžinerijos sričių. Paprastai pagrindinis visų prognozės modelių tikslas – kuo tiksliau iš praeities duomenų ekstrapoliuoti eilutės reikšmes į ateitį. Pagrindiniai plačiai žinomi modeliai yra slenkančio vidurkio ir eksponentinio suglodinimo metodai. Laiko eilučių prognozavimo metodai, naudojantys netiesinius ryšius, paprastai pagrįsti statistiniais regresiniais bei ARIMA modeliais. Pagrindinis trumpų laiko eilučių prognozės modelio tikslas – sukonstruoti proceso modelį ir prognozuoti jo reikšmes ateityje, turint nedidelį kiekį praeities duomenų. Kita vertus, trumpoms laiko eilutėms užtenka vieno žingsnio prognozės, kartais kartu su lokaliais minimumų ir maksimumų įverčiais.

Hankelio matricos laiko eilučių prognozės uždaviniuose panaudotos [4], tuo tarpu algebrinis sekos identifikavimo ir prognozės modelis, panaudojant Hankelio matricų rango sąvoką, pateiktas [2] straipsnyje. Pagrindinis straipsnio tikslas – pagerinti algebrinės sekos prognozę, modifikuojant bazinio fragmento identifikacijos procedūrą, panaudojant vidinį glodinimą. Taip išlaikomas balansas tarp algebrinės sekos identifikuoto bazinio fragmento variabilumo ir slenkančio vidurkio glodinančios prognozės.

1 Sekos rango sąvoka

Tarkime, turime realiųjų skaičių seką S:

$$(x_0, x_1, x_2, \ldots) := (x_k; k \in N_0). \tag{1}$$

Iš šios sekos elementų sukonstruota Hankelio matrica $H^{(n)}$ apibrėžiama taip:

$$H^{(n)} := \begin{bmatrix} x_0 & x_1 & \cdots & x_{n-1} \\ x_1 & x_2 & \cdots & x_n \\ & & \vdots & \\ x_{n-1} & x_n & \cdots & x_{2n-2} \end{bmatrix}, \tag{2}$$

čia n – kvadratinės matricos eilė. Tuomet Hankelio matricos determinantų seką pažymėsime: $d^{(n)} = \det H^{(n)}$, $n \geqslant 1$.

Sekos $(x_k; k \in N_0)$ rangu vadinamas toks natūralusis skaičius $m = Hr(x_k; k \in N_0)$, tenkinantis sąlygą:

$$d^{(m+k)} = 0. \tag{3}$$

visiems $k \in N$, kai $d^{(m)} \neq 0$. Laikykime, kad sekos rangas $Hr(x_k; k \in N_0) = m$, $m < \infty$. Tuomet teisinga ši lygybė [3]:

$$x_n = \sum_{k=1}^{r} \sum_{l=0}^{n_k-1} \mu_{kl} \binom{n}{l} \rho_k^{n-l}, \quad n = 0, 1, 2, \ldots, \tag{4}$$

čia $\rho_k \in C$, $k = 1, 2, \ldots, r$ – charakteristinės šaknys, skačiuojamos iš lygties:

$$\begin{vmatrix} x_0 & x_1 & \cdots & x_m \\ x_1 & x_2 & \cdots & x_{m+1} \\ & & \vdots & \\ x_{m-1} & x_m & \cdots & x_{2m-1} \\ 1 & \rho & \cdots & \rho^m \end{vmatrix} = 0, \tag{5}$$

rekurentiniai šaknų indeksai n_k ($n_k \in N$) tenkina lygtį $n_1 + n_2 + \cdots + n_r = m$, o koeficientai $\mu_{kl} \in C$, $k = 1, 2, \ldots, r$, $l = 0, 1, \ldots, n_k - 1$ randami iš tiesinės lygčių sistemos (4), turinčios tik vieną sprendinį.

2 Prognozės modelis

Tarkime, turime $2n$ stebėjimų: $x_0, x_1, x_2, \ldots, x_{2n-1}$, čia x_{2n-1} – dabarties momento reikšmė. Laikykime, kad sekos rangas lygus n. Tuomet kitas elementas x_{2n} tiesiogiai ir vienareikšmiškai suskaičiuojamas iš lygties:

$$\det H^{(n+1)} = \det \begin{bmatrix} x_0 & x_1 & \cdots & x_n \\ x_1 & x_2 & \cdots & x_{n+1} \\ & & \vdots & \\ x_n & x_{n+1} & \cdots & x_{2n} \end{bmatrix} = 0. \tag{6}$$

Tačiau, šis tiesioginis x_{2n} sekos nario skaičiavimas tinka tik sekoms, turinčioms algebrinį sąryšį. Tačiau, jei (1) seka nėra algebrinė progresija, kyla klausimas, ar įmanoma sukonstruoti modelį, naudojant pirmajame skyriuje pateiktas priklausomybes, nes realaus pasaulio sekos turi didesnį, ar mažesnį triukšmą. Todėl triukšmo pašalinimas ir bazinio algebrinės sekos fragmento identifikavimas tampa aktualiu uždaviniu.

Priimkime prielaidą, kad turime $2n$ stebėjimą ir (1) seka sudaryta iš nežinomos „užtriukšmintos“ algebrinės sekos:

$$x_k = \tilde{x} + \varepsilon_k, \quad k = 0, 1, 2, \ldots, \tag{7}$$

čia ε_k, $k = 0, 1, 2, \ldots$ – triukšmas.

Tarkime, kad $\tilde{x}_k$, $k = 0, 1, 2, \ldots$ yra algebrinė išraiška, t. y. skeletinė seka, apibrėžianti globalią eilutės dinamiką. Tuomet pagal (6) determinantas

$$\det \tilde{H}^{(n+1)} = 0. \tag{8}$$

Kyla klausimas, kaip identifikuoti triukšmą. Toks uždavinys turi be galo daug sprendinių. Mūsų tikslas – minimizuoti bet kokius nuokrypius nuo tariamos algebrinės išraiškos. Kad sušvelnintume pagal (8) lygtį suskaičiuotą prognozę $\tilde{x}_{2n}$, įvedame nuokrypio nuo slenkančio vidurkio prognozės komponentę. Slenkančio vidurkio prognozė skaičiuojama pagal formulę $\bar{x}_k = \frac{1}{s}\sum_{i=0}^{s-1} x_{k-i-1}$, čia s – slenkančio vidurkio prognozės langas. Todėl triukšmų sekai $\{\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{2n-1}\}$ siūlome tokią tikslo funkciją, kurią reikia maksimizuoti:

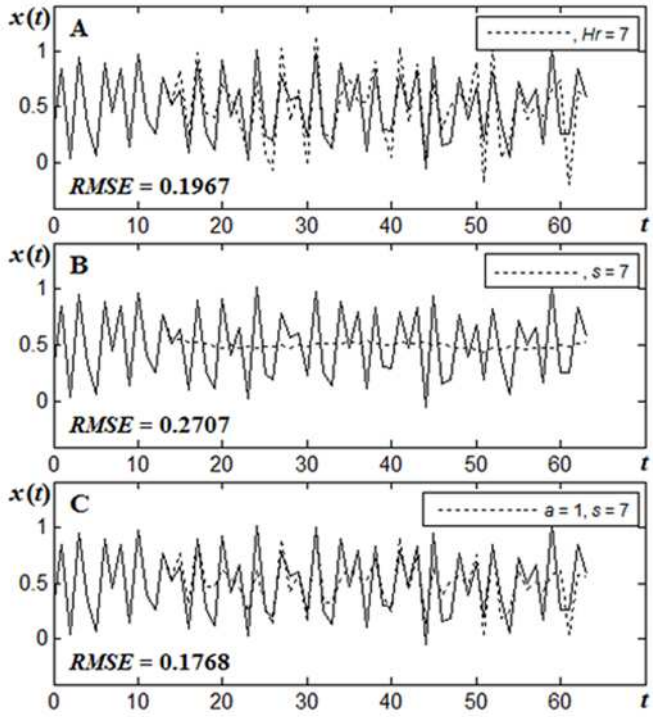
$$F(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{2n-1}) = \frac{1}{a\sum_{k=0}^{2n-1} |\varepsilon_k| + |\tilde{x}_{2n} - \bar{x}_{2n}|}, \tag{9}$$

čia $\tilde{x}_{2n}$ (8) lygties sprendinys, $\bar{x}_{2n}$ – slenkančio vidurkio prognozė, $a > 0$. Siekiame, kad ε_k, $k = 0, 1, 2, \ldots, 2n-1$ būtų kuo mažesni ir algebrinės prognozė nenutoltų nuo slenkančio vidurkio prognozės. Balansas tarp šių dydžių koreguojamas parenkant atitinkamą parametro a reikšmę.

3 Skaitiniai eksperimentai

Kadangi tikslo funkcijai (9) maksimizuoti negalime parinkti tikslaus optimizavimo metodo, tenka naudoti evoliucinius algoritmus ir ieškoti beveik optimalaus sprendinio. Tam tikslui naudojame dalelių spiečiaus algoritmą. Kad galėtume patikrinti modelio funkcionalumą ir parinkti tinkamus parametrus, sugeneruojame dirbtinę periodinę seką, sudarytą iš 7 elementų: 0,5, 0,7, 0,1, 0,9, 0,3, 0,2, 0,8. Ši seka reprezentuoja skeletinę algebrinę seką. Tuomet prie kiekvieno sekos elemento pridedame pagal tolygųjį skirstinį iš intervalo $[-0,15; 0,15]$ pasiskirsčiusius dydžius, taip seka „užtriukšminama“.

Pirmoji užduotis yra identifikuoti H-rangą. Tam panaudojame (8) lygtį. Prognozė atliekama vieną žingsnį į priekį. Tuomet stebėjimų langą perkeliame per vieną reikšmę į priekį ir analogiškai atliekame tiesioginę prognozę. Tokį vieno žingsnio perkėlimą atliekame 50 reikšmių į priekį. Prognozės paklaidas įvertiname panaudodami šaknies iš vidutinės kvadratinės paklaidos (RMSE) metriką. H-rango parinkimas pagal (8) formulę yra jautrus uždavinys. Parinkus per mažą rangą ($Hr = 6$) gaunamos didelės paklaidos, RMSE = 21,5646. Kai rangas per didelis ($Hr = 8$) – situacija panaši, RMSE = 28,7332. Kaip ir buvo galima tikėtis, dirbtinei periodinei eilutei su periodu 7 ir tolygiuoju triukšmu mažiausia paklaida gauta, kai $Hr = 7$, RMSE = 0,1967. Toks H-rango parinkimo modelis (9) tikslo funkcijai bus panaudotas ir realiems duomenims. Nustačius tinkamą rangą, ieškome triukšmų rinkinio pagal (9) funkciją.

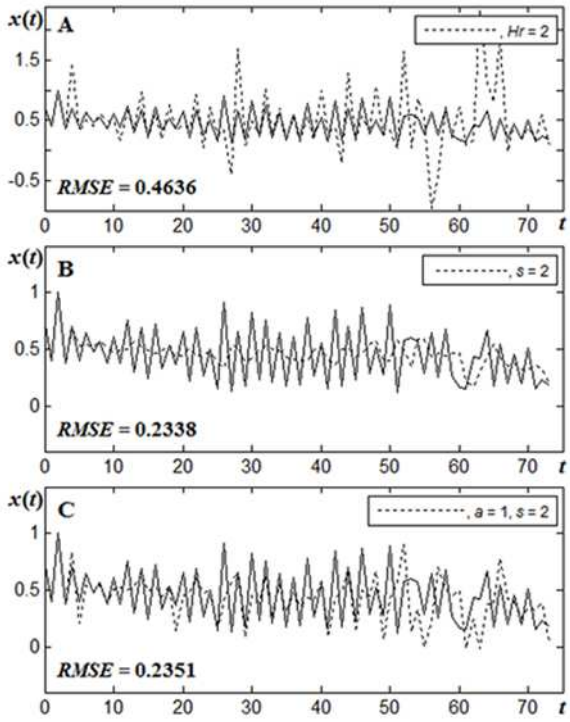


1 pav.
Dirbtinės sekos su tolygiuoju triukšmu prognozės rezultatai (prognozė pavaizduota punktyrine linija): A) tiesioginė algebrinė prognozė, kai $Hr = 7$; B) Slenkančio vidurkio prognozė, kai $s = 7$; C) Metodo, panaudojant vidinį glodinimą prognozė, kai $a = 1$, $s = 7$.

Šiame sprendimo etape svarbu parinkti slenkančio vidurkio langa ilgį s ir parametro a reikšmę (9) tikslo funkcijai. Buvo atlikti skaitiniai eksperimentai renkantis parametro s ir parametro a reikšmes. Geriausi rezultatai pasiekti, kai $a = 1$, o $s = 7$, t. y. į slenkančio vidurkio prognozę įtraukiamų reikšmių skaičius sutampa su nustatytu sekos rangu.

Metodo prognozės rezultatai sekai su tolygiuoju triukšmu pavaizduoti 1 paveiksle. A dalyje pateikta tiesioginė algebrinė prognozė su RMSE = 0,1967, B dalyje slenkančio vidurkio prognozė, kai $s = 7$, RMSE = 0,2707. C dalyje pateiktas bendras metodas, kuomet geriausias rezultatas pasiektas su RMSE = 0,1768.

Kitas eksperimentas atliktas su realiais duomenimis iš internetinės duomenų bazės [1]. Pasirinkta Andrews46 eilutė (1852–1925 m. Broadbalk ūkio surinktų šiaudų kiekis), susidedanti iš 74 narių. Šios eilutės H-rangas parenkamas atliekant tiesioginę prognozę ir geriausi rezultatai gauti, kai $Hr = 2$. Taigi parametrų reikšmės $m = s = 2$. Prognozės rezultatai pateikti 2 paveiksle. Dėl ribotos straipsnio apimties apsiribojame šių dviejų laiko eilučių prognozės pavyzdžiais ir neatliekame statistinės paklaidų analizės, o išvadas apie parametrų parinkimą atliekame remdamiesi paklaidų RMSE įverčiais.



2 pav.
Andrews46 eilutės prognozės rezultatai (prognozė pavaizduota punktyrine linija): A) tiesioginė algebrinė prognozė, kai $Hr = 2$; B) Slenkančio vidurkio prognozė, kai $s = 2$; C) Metodo, panaudojant vidinį glodinimą prognozė, kai $a = 1$, $s = 2$.

4 Išvados

Šiame straipsnyje pateikta trumpų laiko eilučių algebrinės prognozės metodika, identifikuojant skeletinę seką ir ją suglodinant slenkančio vidurkio prognoze. Parenkami tinkami parametrai, kurie išlaiko pusiausvyrą tarp algebrinės kreivės variabilumo ir slenkančio vidurkio prognozės glodinimo, kai prarandama informacija apie sekos įgijamas maksimalias ir minimalias reikšmes. Atlikti skaitiniai eksperimentai su dirbtine seka ir realaus pasaulio eilute parodė, kad geriausi rezultatai pasiekiami, kai pusiausvyra tarp abiejų komponenčių vienoda, tačiau galima tikėtis, kad kitoms sekoms tai parenkama individualiai.

Literatūra

[1] R.J. Hyndman. Time series data library. Available from Internet: http://robjhyndman.com/TSDL/.

[2] V. G. Moskvina and A. A. Zhigljavsky. An algorithm based on singular spectrum analysis for change-point detection. *Comm. Stat. Sim. Comp.*, **32**(2):319–352, 2003.

[3] Z. Navickas and L. Bikulciene. Expressions of solutions of ordinary differential equations by standard functions. *Math. Mod. Anal.*, **75**(1):399–412, 2006.

[4] M. Ragulskis, K. Lukoseviciute, Z. Navickas and R. Palivonaitė. Short-term time series forecasting based on the identification on skeleton algebraic sequences. *Neurocomputing*, **64**:1735–1747, 2011.

SUMMARY

Short-term time series prediction based on sceleton algebraic sequences with internal smoothing

R. Palivonaite, M. Ragulskis

An improved algebraic forecasting method with internal smoothing is proposed for short-term time series prediction. The concept of the H-rank is proposed for the detection of a base fragment of the sequence. Numerical experiments with artificially generated and real-world time series are used to illustrate the forecast method.

Keywords: Hankel matrix, time series forecasting, algebraic sequence.